Accelerated and Sparse Algorithms for Approximate Personalized PageRank and Beyond

David Martínez-Rubio (TUB, ZIB), Elias Wirth (TUB, ZIB), Sebastian Pokutta (TUB, ZIB)

ZIB ZUSE INSTITUTE BERLIN

Cooperation: TU Berlin
Funding: DFG Cluster of Excellence MATH+

Berlin Mathematics Research Center MATH+

Problem [COLT 2022 Open Problem]

Design **local** accelerated methods for ℓ_1-regularized undirected Personalized PageRank problems: depend only on nodes in the solution and neighbors. More generally:

$$\min_{x \in \mathbb{R}^n_{\geq 0}} \{g(\mathbf{x}) \stackrel{\text{def}}{=} \langle \mathbf{x}, Q\mathbf{x}\rangle - \langle \mathbf{b}, \mathbf{x}\rangle\}$$

for symmetric Q s.t. $0 \prec \mu \cdot I \preccurlyeq Q \preccurlyeq L \cdot I$ and $Q_{ij} \leq 0$.

Problem Derivation and Applications

Applications: local graph clustering, i.e,. dividing a graph into internally similar sub-clusters. Used in domains including technical [Vir03; ACL06], biological [XOX02; BH03; BML+05], sociological settings [New03; TMP11], coauthorship networks [LBN+05], etc.

Let G be an undirected connected graph ("dangling nodes" can be dealt with [EMT04]).
The PageRank problem:

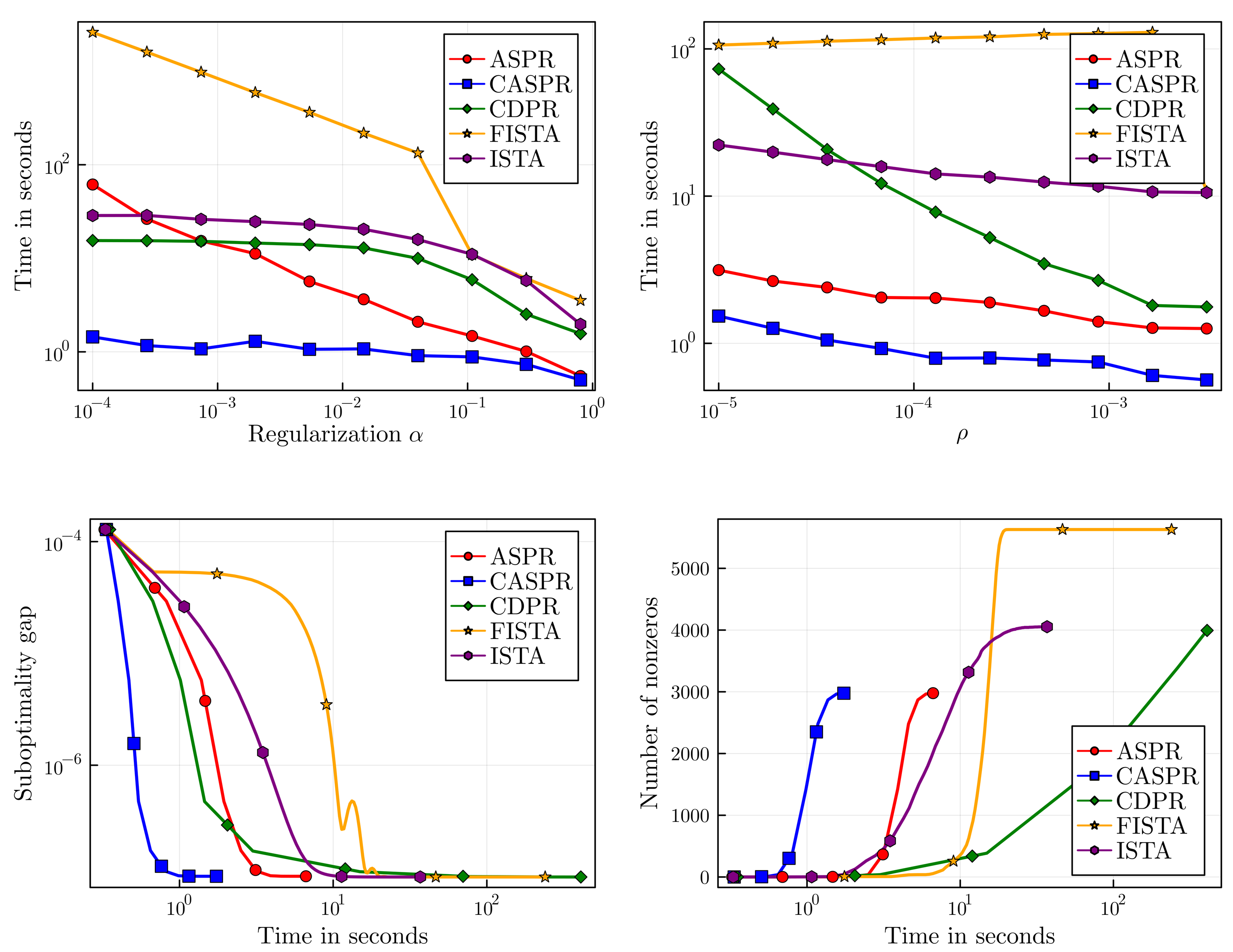
- **Find the stationary distribution** $x \in \Delta$ of the uniform random walk $AD^{-1}x = x$ (A: adjacency matrix, D: diagonal degree matrix). Weighted walk works as well.
- Connectedness $\implies$ irreducibility $\implies$ unique stationary distribution.
- In directed PageRank, **uniqueness** is ensured by modifying the Markov Chain as $(1-\alpha)AD^{-1} + \alpha s \mathbb{1}^T$ for $s \in \Delta$ (**teleportation distribution**) such that we have irreducibility (strong connectedness). In undirected PageRank, this is useful for local graph clusering (e.g. $s = e_i$ computes a cluster around node i).
- Classically in PageRank $s = \mathbb{1}/n$. **Personalized PageRank**: any $s \in \Delta$.
- **Aperiodicity**: We can use the lazy walk $\frac{1}{2}(I + AD^{-1})$ instead of AD^{-1} to ensure convergence of the chain to the stationary distribution.
- Adding ℓ_1-**regularization** $\rho\|x\|_1$ induces sparsity of the solution.
- After reformulation, the problem is in the form above where

$$Q \stackrel{\text{def}}{=} \alpha I + \frac{1-\alpha}{2}\mathcal{L} \quad \text{and} \quad \mathbf{b} \stackrel{\text{def}}{=} \alpha\left(D^{-1/2}\mathbf{s} - \rho D^{1/2}\mathbb{1}\right)$$

  where $\mathcal{L}$ is the symm. normalized Laplacian $I - D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$, which gives $0 \prec \alpha I \preccurlyeq Q \preccurlyeq I$.
- **Beyond PageRank**, we only require symmetric positive definite M-matrix Q.

- We solve the open problem.
- We obtain accelerated sparse algorithms and other sparse algorithms.
- Improvement depends on the sparsity of the solution and the graph.
- Our new algorithms also perform better empirically.

Experiments



Top: Running time along with ISTA and FISTA in a 4M nodes graph. FISTA is accelerated but not guaranteed to be sparse. We fixed $\varepsilon = 10^{-6}$, and either fixed $\rho = 10^{-4}$ while varying α or we fixed $\alpha = 0.05$ and while varying ρ. Bottom: suboptimality gap and number of nonzeroes of the solution with time. **CASPR performs best in practice, as expected from our theory. Our algorithms enjoy better sparsity.** This is due to them sequentially optimizing in the subspace spanned by the current support.

Results and Previous Work

L smoothness, μ strong convexity, $\alpha \leq \mu$ estimate of strong convexity. $\mathcal{S}^* \stackrel{\text{def}}{=} \operatorname{supp}(\mathbf{x}^*)$, $\operatorname{vol}(\mathcal{S}^*) \stackrel{\text{def}}{=} \operatorname{nnz}(Q_{:,\mathcal{S}^*})$ and $\widetilde{\operatorname{vol}}(\mathcal{S}^*) \stackrel{\text{def}}{=} \operatorname{nnz}(Q_{\mathcal{S}^*,\mathcal{S}^*})$.

Method	Time complexity	Space complexity
ISTA [FRS+19]	$\widetilde{\mathcal{O}}(\operatorname{vol}(\mathcal{S}^*)\frac{L}{\mu})$	$\mathcal{O}(\lvert\mathcal{S}^*\rvert)$
CDPR **(Ours)**	$\mathcal{O}(\lvert\mathcal{S}^*\rvert^3 + \lvert\mathcal{S}^*\rvert\operatorname{vol}(\mathcal{S}^*))$	$\mathcal{O}(\lvert\mathcal{S}^*\rvert^2)$
ASPR **(Ours)**	$\widetilde{\mathcal{O}}(\lvert\mathcal{S}^*\rvert\widetilde{\operatorname{vol}}(\mathcal{S}^*)\sqrt{\frac{L}{\alpha}} + \lvert\mathcal{S}^*\rvert\operatorname{vol}(\mathcal{S}^*))$	$\mathcal{O}(\lvert\mathcal{S}^*\rvert)$
CASPR **(Ours)**	$\widetilde{\mathcal{O}}(\lvert\mathcal{S}^*\rvert\widetilde{\operatorname{vol}}(\mathcal{S}^*)\min\left\{\sqrt{\frac{L}{\mu}}, \lvert\mathcal{S}^*\rvert\right\} + \lvert\mathcal{S}^*\rvert\operatorname{vol}(\mathcal{S}^*))$	$\mathcal{O}(\lvert\mathcal{S}^*\rvert)$
LASPR **(Ours)**	$\widetilde{\mathcal{O}}(\lvert\mathcal{S}^*\rvert\operatorname{vol}(\mathcal{S}^*))$	$\mathcal{O}(\lvert\mathcal{S}^*\rvert)$

Exact Algorithmic Scheme

- **Definition:** i is a good coordinate iff $i \in \mathcal{S}^*$. Otherwise it is bad.
- **Idea for an algorithm:** discover good coordinates sequentially, by optimizing in $C^{(t)} \stackrel{\text{def}}{=} \operatorname{span}(\{\mathbf{e}_i \mid i \in S^{(t)}\}) \cap \mathbb{R}^n_{\geq 0}$, where $S^{(t)}$ is the set of known good coordinates.
- By the geometric lemma below, at the minimizer $\mathbf{x}^{(*,t+1)} \stackrel{\text{def}}{=} \mathbf{x}^{(*,C^{(t)})}$ we have $\nabla_i g(\mathbf{x}^{(*,t+1)}) < 0$ only if i is good and new, i.e., only if $i \in \mathcal{S}^* \setminus S^{(t)}$.
- Using conjugate directions and exploiting the sparsity we define CDPR.

Geometric Lemma

For $S \subseteq [n]$ and $\mathbf{x}^{(t)} \in \mathbb{R}^n$ such that $x_i = 0$ if $i \in S$ and $\nabla_i g(\mathbf{x}) \leq 0$ if $i \in S$. Let $C \stackrel{\text{def}}{=} \operatorname{span}(\{\mathbf{e}_i \mid i \in S\}) \cap \mathbb{R}^n_{\geq 0}$, $\mathbf{x}^{(*,C)} \stackrel{\text{def}}{=} \arg\min_{\mathbf{x} \in C} g(\mathbf{x})$, $\mathbf{x}^* \stackrel{\text{def}}{=} \arg\min_{\mathbf{x} \in \mathbb{R}^n_{\geq 0}} g(\mathbf{x})$.

1. It holds that $\mathbf{x}^{(0)} \leq \mathbf{x}^{(*,C)}$ and $\nabla_i g(\mathbf{x}^{(*,C)}) = 0$ for all $i \in S$.
2. If for $i \in S$, we have $\mathbf{x}_i^{(0)} > 0$ or $\nabla_i g(\mathbf{x}^{(0)}) < 0$, then $x_i^{(*,C)} > 0$.
3. If $x_i^{(*,C)} > 0$ for all $i \in S$, we have $\mathbf{x}^{(*,C)} \leq \mathbf{x}^*$ and therefore $S \subseteq \mathcal{S}^*$.

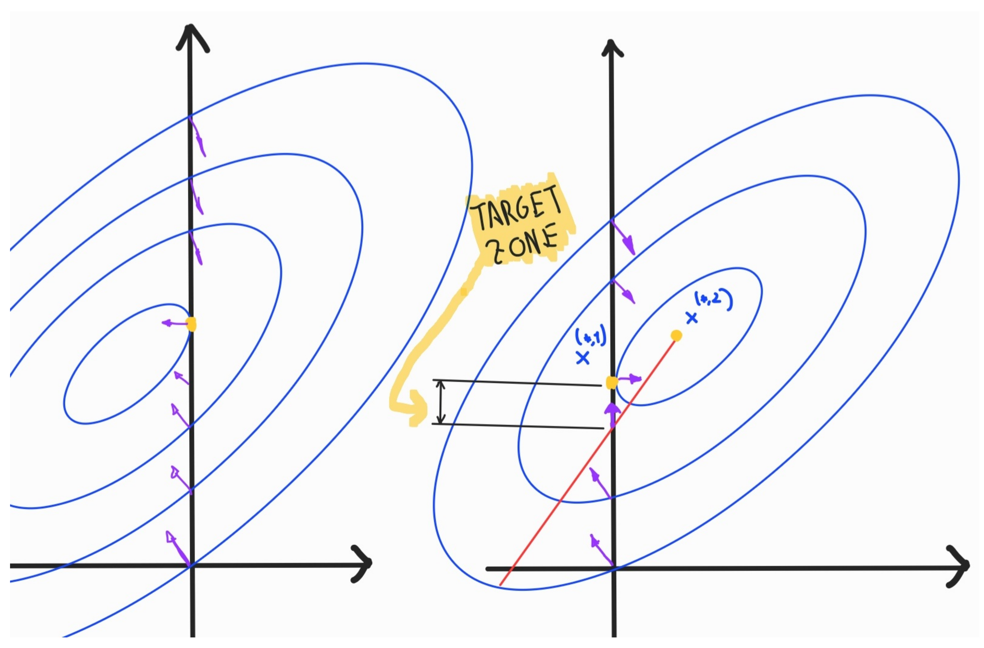


Fig. 3: **Right:** A negative coordinate gradient like $\nabla_2 g(\mathbf{x})$ for a point below the optimizer in $\operatorname{span}(e_1)$ implies the coordinate is good. So we search for points with $\nabla_S g(\mathbf{x}) \leq 0$, s.t. $\exists i \notin S, \nabla_i g(\mathbf{x}) < 0$, for $S = \operatorname{supp}(\mathbf{x})$. **Left:** A negative coordinate gradient like $\nabla_2 g(\mathbf{x})$ for a point not below the optimizer in $\operatorname{span}(e_1)$ does not imply the coordinate is good.

Proof of 1.: $\bar{g} \stackrel{\text{def}}{=} g$ restricted to C. Let $\{\mathbf{x}^{(t)}\}_{t=0}^{\infty}$ be the iterates of GD$(C, \mathbf{x}^{(0)}, \bar{g})$. By induction, $\nabla\bar{g}(\mathbf{x}^{(t)}) \leq 0$ and $\mathbf{x}^{(t)} \leq \mathbf{x}^{(t+1)} \in C$. Indeed, using that the gradient is affine:

$$\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} - \underbrace{{}^1\!/\!_L\nabla\bar{g}(\mathbf{x}^{(t)})}_{\leq 0} \geq \mathbf{x}^{(t)} \text{ and } \nabla\bar{g}(\mathbf{x}^{(t+1)}) = \underbrace{\nabla\bar{g}(\mathbf{x}^{(t)})}_{\leq 0}\underbrace{(I - {}^1\!/\!_L Q_{S,S})}_{\geq 0} \leq 0$$

$\mathbf{x}^{(t)} \to \mathbf{x}^{(*,C)}$, $\nabla\bar{g}(\mathbf{x}^{(t)}) \to \nabla\bar{g}(\mathbf{x}^{(*,C)}) \leq 0$ (and by optimality ≥ 0.)

Approximate Accelerated Algorithmic Scheme

1. Because $Q_{ij} \leq 0$ for $i \neq j$, for $\mathbf{y} = \mathbf{x} - \Delta\mathbf{e}_i$, we have $\forall j \neq i$: $\nabla_j g(\mathbf{y}) \geq \nabla_j g(\mathbf{x})$ if $\Delta > 0$.
2. Recall, $\nabla_i g(\mathbf{x}^{(*,C^{(t)})}) < 0$ only if i is good. So by 1., for $\mathbf{x} \in C^{(t)}$ s.t. $\mathbf{x} \leq \mathbf{x}^{(*,C^{(t)})}$, new coordinates i can only satisfy $\nabla_i g(\mathbf{x}) < 0$ if they are good.
3. **Strategy**: Get close to $\mathbf{x}^{(*,C^{(t)})}$ and then move slightly towards 0 to be $\leq \mathbf{x}^{(*,C^{(t)})}$.
4. **Lemma**. Let $\bar{\mathbf{x}}^{(t+1)}$ be an $\varepsilon \cdot \frac{\alpha^2}{2(1+|S^{(t)}|)L^2}$ minimizer in $C^{(t)}$. Define $\mathbf{x}^{(t+1)} \leftarrow \operatorname{Proj}_{\mathbb{R}^n_{\geq 0}}(\bar{\mathbf{x}}^{(t+1)} - \delta_t \mathbb{1})$ for $\delta_t = \sqrt{\frac{\varepsilon\alpha}{(1+|S^{(t)}|)L^2}}$. Then, $\mathbf{x}^{(t+1)} \leq \mathbf{x}^{(*,C^{(t)})}$ and $\mathbf{x}^{(t+1)}$ is a global ε-minimizer or there is i s.t. $\nabla_i g(\mathbf{x}^{(t+1)}) < 0$, so we expand the current set of good coordinates $S^{(t)}$.
5. **Intuition**. $\mathbf{x}^{(t+1)}$ is almost optimal in $C^{(t)}$, so if its global gap is $> \varepsilon$ then 1 step of GD makes more progress than what it is possible in $C^{(t)}$. $\implies \exists i \notin S^{(t)}$ s.t. $\nabla_i g(\mathbf{x}^{(t+1)}) < 0$.
6. **Obtain 3.** with projected AGD or Conjugate Gradients with the right stopping criterion (our problem is strongly convex and smooth). We can also use a nearly-linear time solver for symmetric diagonally dominant linear systems.